METHOD AND SYSTEM FOR UMTS PACKET TRANSMISSION
SCHEDULING ON SHARED DOWNLINK CHANNELS

Cross Reference To Related Application

This application claims priority of European Application No. 00310343.9
5   filed on November 22, 2000.

BACKGROUND OF THE INVENTION

Field of the invention

The invention relates in general to a method for packet transmission scheduling
and to a packet transmission scheduling system and specifically to a UMTS
10 packet transmission scheduling method and to a System having UMTS packet
transmission scheduling functionality.

Description of the Related Art

The invention addresses the problem of efficient scheduling of data flows,
15 especially in Universal Mobile Telecommunication Systems (UMTS) for a transfer
of data between a local base transceiver station (BTS) and e.g. a mobile station
also termed user equipment (UE), i.e. for the UMTS Downlink Shared Channel
(DSCH) between the base transceiver station and the mobile station. In the base
transceiver station the radio access network has a complete or perfect
knowledge of how much, when, and to whom packet transmissions are made or
20 will have to be made within a certain interval of time. Hence, a central controlled
unit may be and is used for the downlink data transmission.

However, in packet switching networks the task of multiplexing essentially
25 reduces to the task of ordering packets and to then send them serially over a
shared link. This process of serialization is referred to as scheduling. The benefit
of packet switching is based on the multiplexing gain, where some data flows
benefit from unused resources from other temporarily inactive data flows. The

disadvantage of this service compared to a circuit switched service is a lack of predictability in such systems. Significantly, predictability of system behavior is one important measure of quality. Some services as e.g. internet communication or facsimile transmission services need stronger Quality of Service or „QoS"

5   guarantees than others as for example pure speech data transmission. Quality of service (QoS) scheduling tries to balance the amount and timing of the received service for each flow according to the respective data flow requirements.

SUMMARY OF THE INVENTION

10   It is an object of the invention to provide an improved method for packet transmission scheduling, especially on downlink shared channels and an improved packet transmission scheduling system, both the improved method and the system especially adapted to be used for UMTS systems.

15   The inventive solution is characterized by a method, a system, a base and mobile station and an implementation software incorporating the features of claim 1, 15, 24, 25 and 26, respectively.

Accordingly, the invention proposes and uses a QoS-scheduling for handling
20   multiple data flows in a Code Division Multiple Access (CDMA) system by dynamically scheduling protocol data units (PDU) in dependence of allocated radio resource constraints, especially ensuring the required data rates due to a rate conserving scheduling by simultaneously performing an optimization of the usage of radio resources.

25
The inventive QoS-scheduling preferably is apt to handle the data flows on downlink shared channels but can also be applied to scheduling of multiple data flows for different users on a dedicated channel in the downlink direction and for a single user in the uplink direction.

According to the preferred embodiments, the invention relies on two schedulers, which are linked together in a novel manner, whereby the first scheduler provides a certain degree of predictable behavior and the second scheduler provides Medium Access Control (MAC access) and also allows for bandwidth conserving segmentation and allocation strategies.

These two schedulers are named PDU scheduler and MAC scheduler. For the basics of this scheduling method reference is made to the co-pending application filed with the same Patent Office, Stefan Gruhl: "Method of linking two schedulers of a multi-layer network and a network comprising a transceiver having linking functionality for two schedulers", the contents of which is incorporated by reference hereby. In the present application proposal the scheduling method of the above-cited co-pending application has been adapted to the UMTS mobile communication system. It is especially shown how to link the two schedulers and parameterize the local algorithms for UMTS, especially for the medium access control (MAC) scheduler.

BRIEF DESCRIPTION OF THE DRAWINGS

The invention is described in view of preferred embodiments in more detail below and reference is made to the accompanying drawings.

Figure 1: shows a message flow for adding a radio
bearer or user equipment (UE) to the scheduler;

Figure 2: shows the allocation of channelisation codes using
the code branch allocation (CBA) method in view of a preferred
embodiment;

Figure 3: shows principles of the quality of service (QoS)
scheduling method;

Figure 4: shows an improved medium access control (MAC)

                scheduling mechanism;

Figure 5: shows how to process power limits within an

                improved medium access (MAC) scheduler based on a preferred

                embodiment.

DETAILED DESCRIPTION

For a better understanding of the invention and especially to ensure the
improved performance of the inventive method and devices, certain
requirements should be met and a number of assumptions are made in advance.

Assumptions and Requirements

A certain amount $P_{PS} = \alpha_{PS} \cdot P_{max}$ of the overall maximum transmission power P_{max}
is allocated by the radio resource management unit (RRM) to the packet
switched radio bearer. The scheduler is apt to use P_{PS} essentially autonomously
without invocation of the radio resource management unit (RRM).

If automatic repeat request (ARQ) is applied, the number of retransmissions is
assumed to be significantly smaller than the regular traffic.

All transmissions having certain assigned quality requirements are embedded
into a data flow. Accordingly, a data flow is defined as a sequence of data
packets from the same source to the same destination in the network, for which
the user has certain QoS requirements.

Each radio bearer is related to a single data flow. Because multiple radio bearers
might be established for a single user, multiple data flows could exist
simultaneously that are related to a single user, too. In the following description
all data flows are handled separately.

Throughout this description the elements of a data flow are defined as Protocol Data Units (PDU or PDUs).

5   These PDU are typically layer 3 elements from an UMTS point of view, but the invention is not necessarily limited to this.

Protocol data units (PDUs) are segmented into transport blocks (TB or TBs), which receive their own headers, as specified in the UMTS 3GPP standards.
10   This operation is associated with layer-2. Typically but not necessarily the transport blocks have a fixed size. An arbitrary number of transport blocks can be put together to one Transport Block Set (TBS). Typically but not necessarily only transport blocks of one protocol data unit (PDU) are put together.

15   One TBS is scheduled by the medium access control layer within the scheduling interval (typically 10ms) to the physical layer (PHY-layer) per flow.

For the downlink shared channel there is no soft handover (HO) assumed. Therefore, the scheduler handles the data flows addressed to UEs of the own
20   cell, only.

Any mobility related procedures, e.g. hard handover are handled by the radio resource management system (RRM) independently.

25   The Bit Error Rate (BER) of a data flow is a static QoS requirement of the associated radio bearer. Depending on the delay constraints, there is a trade-off between Forward Error Correction (FEC), i.e. the received signal energy to noise ratio E_b/N_0 vs. automatic repeat request (ARQ) methods, i.e. the allowed number of retransmissions.

It is assumed that the required bit error rate always can be received from the core network or radio access network.

5   In order to optimize the bandwidth consumption of a data flow padding is minimized as a trade-off for delay where possible. This is indicated by the flow's QoS constraints and the recent flow status.

In order to optimize the delay, the whole protocol data unit (PDU) is taken by the
10   PDU scheduler preferably at once.

The downlink shared channel preferably is time synchronized, i.e. every data flow starts its transmission at the same point of time. Discontinuous transmission (DTX) on the downlink shared channel could lead to large fluctuations in
15   interference to users on other dedicated channels (DCH) of the same cell or to all users of the adjacent cells. Hence, a discontinuous transmission (DTX) should not be used in downlink shared channels.

According to recent 3GPP standards there is no physical multiplexing (or PHY
20   MuX) for different data flows in the downlink shared channel. As a consequence thereof, the transport format combination sets (TFCS) on the downlink shared channel consists of a transport format set (TFS) for one data flow, only. The transport format set is associated to the data rates R_B of the respective data flow. The transport format sets are directly related to the spreading factor SF of the
25   Code Division Multiple Access (CDMA) transmission system which is used to support that data rate.

The transport block size within one scheduling interval remains constant for each protocol data unit (PDU). Consequently, only the number of transport blocks

needs to be counted for medium access control (MAC) scheduling.

Radio Resource Allocation (RRA)

5   Basics on Radio Resource Allocation for a downlink shared channel
Due to its nature and in view of using a Code Division Multiple Access (CDMA)
method, the main resource in the UMTS mobile communication system is the
transmission power, which has to be spent for a certain user. The transmission
power P_{tri} of data flow #i is expressed as

$$P_{tri} \approx \left(\frac{E_B}{N_0}\right)_i \cdot \frac{R_{Bi}}{W} \cdot \frac{I_{0i}}{h_i} = R_{Bi} \cdot C_i \qquad \textbf{equation 1}$$

where
$(E_B/N_0)_i$ denotes the to be received signal energy to noise
15   ratio for data flow # i,
R_{Bi} denotes the current data rate used by flow #i,
W is the chip rate which chip rate at the moment is   defined for UMTS as $W =$
3.84 MChip/s,
I_{0i} denotes the interference at the user equipment (UE)   where the downlink
20   data flow is addressed to,
    h_i is the path-loss between the user equipment (UE) and Node B.

Resources in CDMA systems depend not only on the data rate as handled by
conventional scheduling methods or prior art schedulers but also on a factor C
25   which depends on several other parameters, like path-losses and interference.
From equation 1 C_i is given as

$$C_i = \left(\frac{E_B}{N_0}\right)_i \cdot \frac{1}{W} \cdot \frac{I_{O_i}}{h_i} \qquad \textbf{equation 2}$$

The C_i-value is an essential part of the scheduling algorithm. Depending on the
time when it is used, this value preferably is calculated in two ways:

At or during the establishment of the data flow equation 2 may be used directly,
where $(E_B/N_0)_i$ is initially determined from the bit error (BER) requirement of the
respective radio bearer and (I_o/h_i) is estimated from a measurement that is
signaled from the user equipment (UE) to the network.

After a communication channel is established with the radio bearer, there are no
regular measurements of (I_o/h_i) available in the network. Additionally, $(E_B/N_0)_i$
may differ from the initial value, e.g. due to a varying environment. Then,
equation 2 should not used then, i.e. after establishing of the communication
channel. At this moment equation 1 is used with

$$C_i = P_{tri}^{prev} / R_{Bi}^{prev}, \qquad \textbf{equation 3}$$

where P_{tri}^{prev} and R_{Bi}^{prev} are the previous transmission power and previous data
rate of data flow #i, respectively.

The overall transmission power $\sum_{i \in active} P_{tri}$ of all active data flows on the
downlink shared channel is limited by the allocated transmission power P_{PS} for
the packet switched users. Therefore, the overall transmission limit for all active
data flows is

$$\sum_{i \in active} P_{tri} = \sum_{i \in active} R_{Bi} \cdot C_i \le P_{PS} \, . \quad \textbf{equation 4}$$

Due to limited capabilities of e.g. a single channel transmission unit there is a restriction of the transmission power of a single data flow P_{max}^{single}, too.

5 Consequently, in addition to equation 4 the following limit is given for all active data flows

$$P_{tri} = R_{Bi} \cdot C_i \le P_{max}^{single}, \quad \forall i \in active \, . \quad \textbf{equation 5}$$

10 Tasks and Functions of the improved Radio Resource Allocation (RRA)

Figure 1 shows a message flow between a Core Network (CN), a Radio Network Controller (RNC) and User Equipment (UE) when a new radio bearer is added to the scheduling function.

15 Although it is an essential part of a UMTS mobile communication system, the base transceiver station (BTS) is not separately shown as these stations are well known to persons skilled in the art. The following tasks and functions should be performed by the radio resource allocation units (RRA) before the scheduling function or operation starts:

20

1. Radio Bearer Establishment Request

During this phase a new radio bearer establishment is requested from the Core Network. This request must contain or specify the quality of service (QoS) requirements of the associated data flow, i.e. the requested bit error rates (BER),

25 data rates to be transmitted and delay requirements. If there is no radio resource control (RRC) connection established, a radio resource control connection

establishment procedure between the radio network controller (RNC) and the user equipment (UE) has to be performed which is shown as step 1a in Figure 1.

2. Admission Control (AC)

5   The purpose of admission control is to decide whether this new request can be admitted or not. For admission control several parameters like the requested quality (QoS) of the service and the current network load are used. Other reasons for not admitting the request might be that no radio resources are available which is checked by the next step. If the request is denied there might

10   be a negotiation procedure with a lower quality of service QoS.

3. Dynamic Channel Allocation

The Dynamic Channel Allocation (DCA) procedure allocates the following transmission parameters to the data flow (non exclusive): Transport format set (TFS), radio link control (RLC) Info, new Channelisation Code, Initial

15   transmission power etc. For the allocation method for transport format set and channelisation code see also paragraph „Allocation of Data Rates" of this description. A new amount of transmission power P_{packet} for the scheduler can also be allocated by DCA.

20

4. Radio Bearer Setup

This function performs a setup of the Radio Bearer and synchronization between RNC and UE. Additionally, the BTS will be initialized with the parameters that are allocated by DCA which is not shown in Figure 1 but known to a person skilled in

25   the art.

5. Start Dynamic Scheduling

After a successful establishment and initialization the new data flow is added to
the scheduling function. The scheduling function will be performed now also for
this flow. Reference is made to Figure 1 showing the message flow for adding a
5 radio bearer to the scheduler.

Allocation of Data Rates, TFS and Channelisation Codes

Allocation of Data Rates

The allocation of the data rates for each data flow has a strong impact on the
system efficiency that the scheduler can achieve. The data rates relate to TFS
10 and the channelisation codes. According to the recent 3GPP standards there is
no physical multiplexing (or PHY MuX) for different data flows in the downlink
shared channel. As a consequence thereof, the transport format combination
sets (TFCS) on the downlink shared channel consists of a transport format set
(TFS) for one data flow, only. The transport format set is associated to the data
15 rates R_B of the respective data flow. The transport format sets are directly related
to the spreading factor SF of the Code Division Multiple Access (CDMA)
transmission system which is used to support that data rate.
For a rough allocation or estimation of the limit data rates the following rules are
applied:

20

For the maximum data rate R_{Bmax} transport format sets (TFS) should be allocated
to allow data rates up to two to four times of the maximum data rate, i.e. (2 to
4)R_{Bmax}. There are two reasons for this requirement. The first is that these
maximum transport format sets are required from the MAC-scheduler to serve a
25 flow with a temporarily higher data rate than requested to let a flow profit from
other flows being idle. This applies when there is remaining capacity on the air-
link and if this flow has already sent packets to the PDU list ahead of its specified

rate.

The second reason is to allow for a time divisional multiplex style multiplexing on the transport block level.

The MAC-scheduler's algorithm can be developed towards bandwidth efficiency. Therefore it also may be desirable to use arbitrary sizes of transport blocks to minimise padding. This means the available transport format should be able to temporarily exceed specified rates.

Fairness, bandwidth and quality of service (BW-QoS) guarantees among flows are maintained by another scheduler, the PDU scheduler.

For the minimum data rates R_{Bmin} transport format sets should be allocated to allow data rates below R_{Bmin}. The availability of smaller transport format sets allows the MAC-scheduler to minimise padding. As this introduces additional delay and lowers the average transmission rate, it is only applicable to certain QoS flows. The optimization of such allocation and TFC usage is addressed separately.

In order to comply with the assumption that the downlink shared channel preferably is time synchronized, i.e. every data flow starts its transmission at the same point of time, and discontinuous transmission (DTX) on the downlink shared channel could lead to large fluctuations in interference to users on other dedicated channels (DCH) of the same cell or to all users of the adjacent cells, only those transport format sets are allowed which would fill up the whole data frame with data. Because the spreading factors of the channelisation codes are in the order of $SF = 2^k$, $k = 2,3,...$, according to the aforementioned recent 3GPP standards, this leads to data rates of $R_B = R'_B \cdot 2^n$, $n = 0,1,...$, where R'_B

represents a reference data rate for a certain given spreading factor and may become R_{Bmn}.

Transport Format Set (TFS)

5   The transport format set is defined as the set of transport formats that are associated to one data flow. The semi-static part (coding, transmission interval, rate matching) essentially determines the bit error rate. It is defined by the radio resource management. In the following discussion, the focus is only on the dynamic part which consists of the transport block size and the transport block

10   set size. The dynamic part of transport format sets can be used for optimization of the segmentation in the RLC. For the choice of this dynamic part there is a trade-off between the granularity of the data rates and the limited size of transport format sets. On the one hand, each data flow intends to have a high granularity in data rates avoiding extensive padding. This would lead to a large

15   transport format set size. On the other hand, a transport format set is used to enable efficient physical or PHY signaling for changing data rates. Due to limitations of that PHY signaling (e.g. TFCI (Transport Format Combination Indicator) encoding) the maximum transport format set size is quite limited. Therefore the following transport format set allocation rules regarding the

20   characteristic of the data flow are proposed and used according to the invention:

1. Real Time (RT) Services: This service type needs immediately serving of the offered data. Hence, a high granularity towards higher data rates is desired. Therefore, for real time services a larger transport format set should be

25   allocated.

2. Non Real Time (NRT) Delay Sensitive Services: Here, a limited automatic repeat request (ARQ) can be used for protection of the data flow. The granularity is not as high as for the pure RT service because some data can

be queued for a limited time. Therefore a limited transport format set can be
allocated for such services. For efficient use of automatic repeat request
mechanisms the transport block size should be small.

5   3. NRT Unconstrained Delay Services: This type of service is the best candidate
for bandwidth optimization. In principle unlimited queuing is possible. Hence,
no much granularity is necessary. Therefore a quite limited transport format
set can be allocated for this service type. Granularity is used to avoid
padding, only.

10

Beyond the delay constraint it is sensible to take further QoS requirements and
flow specifications into account. The system could potentially adopt to certain
preferred PDU sizes, like that of a transport control protocol acknowledgement
(TCP-ACK). Bulk data transfer could guard the selection towards max. PDU size,
15   etc.

DL Channelisation Codes
For allocation of downlink (DL) channelisation codes the method of Code Branch
Allocation (CBA) as described in Qiang Cao, Seau Lim, Jens Mueckenheim:
20   „Code branch allocation for CDMA systems," Patent Appl. EP 99 301 810.0, filed
10.03.99 is used. The CBA method is an attempt to solve the code space
shortage problem, which is especially relevant in the downlink. CBA defines a
path in the code tree, the code branch, which consists of spreading codes for
each SF that can be used for the transmission. The code branch might be
25   transmitted to the UE. There can be intersections of code branches where only
exclusive use of one code branch simultaneously is allowed. Because there is a
fixed relation between data rate and spreading factor (SF), see also the above
paragraph assumptions and requirements, an allocation rule of the path in the
code tree is used in view of recent 3GPP standards. According to recent 3GPP

standards there is no physical multiplexing (or PHY MuX) for different data flows
in the downlink shared channel. As a consequence thereof, the transport format
combination sets (TFCS) on the downlink shared channel consists of a transport
format set (TFS) for one data flow, only. The transport format set is associated to

5   the data rates R_B of the respective data flow. The transport format sets are
directly related to the spreading factor SF of the Code Division Multiple Access
(CDMA) transmission system which is used to support that data rate. The
following allocation rule is used according to the invention:

10 1. For data rates below the maximum requested one the nodes of the code
branch should be allocated always below intersection points. This allows
guaranteed data rates R_{Bmax} for all data flows, if no padding-reduction strategy
is used in the MAC-scheduler. For flows that have stringent BW-QoS
requirements and where still padding-reduction is desirable, CBA should only

15   be made with unconstrained flows.

2. For TFC for higher data rates of 2 to 4 times· R_{Bmax} that can be used for the
proactive scheduling (cf. data rate allocation rule 11) nodes in the code tree
can be used that are above intersection points. Because only a quite limited

20   number of data flows is allowed to use this larger data rates no conflicts in the
channelisation code usage are expected.

In the following usage of the CBA method for allocation of channelisation codes
for the scheduling on downlink shared channel is explained by means of an

25 example. Figure 2 illustrates a code tree, which is a representation of the
orthogonal variable spreading factor codes that are used on the UMTS downlink
shared channel. Every node characterizes a code sequence with a spreading
factor given by the first number. All codes within the code tree cannot be used
simultaneously. A node can be used for a physical channel if and only if no other

node on the path from the specific node to the root of the tree (i.e. with lower SF) or in the sub tree below the specific node (i.e. with higher SF) is used by another physical channel. In the example it is assumed that the sub tree below node 4,1 is reserved for downlink shared channel usage. The assignment of the nodes for
5   the code branches to two users can be as follows:

Because the nodes below node 8,1 and 8,2 can be used simultaneously, these sub trees can be·assigned for the data rates below R_{Bmax}. For example, user #1 gets the nodes 8,1 and 16,1. User two will be assigned to nodes 8,2 and 16,4.
10

For the proactive scheduling purpose node 4,1 is allocated to both, user #1 and user #2. They cannot be used simultaneously. Hence, the scheduler must ensure that when user #1 takes node 4,1 user #2 should not transmit on any node of its code branch and vice versa.
15

Hence, the code branch for user #1 is: (4,1); (8,1); (16,1) and for user #2: (4,1); (8,2); (16,4), see also Figure 2.

The Scheduling Method

The invention contemplates using of two schedulers, which are linked together in
20   a novel manner to achieve a certain degree of predictable behavior, while also allowing for bandwidth conserving segmentation and scheduling, see also Stefan Gruhl in "Method of linking two schedulers of a multi-layer network and a network comprising a transceiver having linking functionality for two schedulers". These two schedulers are named PDU scheduler and MAC-scheduler.

25

The first one operates on the input data from Layer 3, the Protocol Data Units (PDU). It receives the QoS requirements of each flows and determines the order in which PDUs should receive service. This service is delivered by lower layer

protocol functionality and incorporates mainly two steps. The first is Layer-2 protocol related and includes mainly segmentation of PDU to Transport Blocks (TB) and Automatic Repeat Request (ARQ).

5   The MAC-scheduler serves the PDU's from this list and tries to reflect the order in the list, while also taking CBA-, timing and power constraints into account.

In Figure 3 the principle architecture of the two serial schedulers is shown. The MAC-scheduler is active at every frame, e.g. on a 10 ms base. The PDU
10   scheduler is operated on all active flows, i.e. with a non-empty PDU flow-queue. Systems of serially uncoupled schedulers can show undesirable behavior. Therefore, both schedulers are linked together.

To allow for the linking the intermediate protocol functionality (depicted as cloud
15   in Figure 3) has to allow for a certain implementation summarized below. It is assumed that layer-2 protocol functionality is apt to be processed as a stateless operation with the character of a function with negligible processing time. Two main tasks in layer-2 are identified which are for the sender, the segmentation process to transport blocks TBs and an ARQ stage for transport blocks TBs. The
20   segmentation process is intuitively a stateless function as it turns one input element in an output vector of new elements (typically with new headers for sequence numbers, etc...). Automatic repeat request ARQ is processed as a function with the input of a TB and the output of the same TB, with ARQ
·   parameters set. Besides this output there is potentially other asynchronous
25   output of the ARQ stage:

a) If the ARQ window size is reached, no output is generated until former transmissions were successfully acknowledged by the receiver. This case can be neglected if always a test on the available window size is done before performing

the ARQ operation.

b) When the automatic repeat request (ARQ) stage receives the request for a retransmission, it generates an output without input.

5

The number of retransmissions is assumed to be significantly smaller than the regular traffic and therefore treat this relatively rare case as a separate ARQ-process which generates its own traffic regardless of the traffic which is mainly running through the ARQ stage in the describe function manner.

10

c) If the maximum number of retransmissions is exceeded and the ARQ considers a PDU transmission unsuccessful, there is another asynchronous signal that has to be signaled to upper layers.

15   The central idea to allow for such linking of the two schedulers is that the MAC-scheduler operates on PDU's in the Protocol Data Unit (PDU) list, while this list is dynamically changed by the Protocol Data Unit (PDU) scheduler. As a consequence, the whole PDU access has to take place via reference to allow for protocol operation on demand techniques. This has to be accompanied by
20   locking of the shared element, the PDU list.

The exact linking method and further details are subject of the above-cited application from Stefan Gruhl: "Method of linking two schedulers of a multi-layer network and a network comprising a transceiver having linking functionality for
25   two schedulers".

PDU Scheduler
The PDU scheduler takes the PDU from the incoming data flows. Each data flow is queued in its own FIFO-queue, denoted as PDU flow-queue. They are

scheduled regarding their QoS requirements into one common list for the MAC-scheduler. This list is denoted as PDU list. This list is not termed a queue, because due to MAC-constraints it cannot be assured to serve this queue in a FIFO fashion. The PDU scheduler must be able to serve the data with the

5   required data rates. For this purpose any rate conserving scheduling policy can be applied, see Hui Zhang, "Service Disciplines for guaranteed Performance Service in Packet-Switching Networks," Proceedings of the IEEE, Vol. 83, No. 10. October 1995, e.g. Weighted Fair Queuing (WF^{2}Q) or Virtual Clock Queuing (VCQ).

10   The scheduling elements for this scheduler are taken depending on the following rule: Regularly the PDU's are large enough to be scheduled as one unit. In this case one scheduling element equals one PDU.

15   If it is possible to serve several PDU's simultaneously on the MAC-layer it may become desirable to have several PDU's available for MAC-scheduling from one flow. This would mainly be the case where PDU's become too small to be served with the required minimum data rate R_{Bmin} within the given MAC-scheduling interval $T_{schedule}$ (typically $T_{schedule}$ = 10ms), i.e. when

20

$$PDU_{len} / R_{B\min} \leq T_{schedule} \qquad \textbf{equation 6}$$

The problem can be overcome by having several PDU's from the flow grouped together into one container, which then becomes the scheduling element.

25

Hence, the scheduling element can be defined as one container that may consist of one (normally) or several PDU's. Throughout this document one scheduling element is defined as PDU and the term PDU scheduler is used for the sake of

simplicity.

As shown in J. Cobb, M. Gouda and A-El-Nahas, "Flow timestamps," Annual
Joint Conference of Information Sciences, 1995, it is also reasonable to work
5   with flow time-stamps instead of PDU time-stamps. When doing so, the inventive
PDU scheduler becomes active when a PDU from a flow is fully served and
therefore removed from the PDU list or when a formerly inactive flow gets
reactivated by a PDU arrival into its empty PDU flow-queue. This is beneficial as
it limits the number of elements in our PDU list to the number of active flows.

10   MAC-Scheduler
Main Function of the MAC Scheduler
The MAC-scheduler serves the PDU's from the PDU scheduler. The order in the
PDU scheduler's list signals the priority in which the PDU scheduler wants the
15   PDU's to be served. The MAC scheduler tries to achieve this while obeying three
restrictions:
Bandwidth constraint due to TFC allocation of the flow and availability of a node
in the code tree (refers to CBA)
Delay constraints:
20   Drives the decision how many subsequent TBS transmissions spread over
several timing intervals are tolerable to obey timing requirements of the served
PDU
ARQ-constraints:
Transmission of TB's that receive ARQ service is only possible until the ARQ
25   window size is reached. Further transmissions are possible only after the ARQ
stage receives the acknowledgement from the receiver.
    Power constraints:
Both the power for the transmission to one individual mobile and the overall
power in the cell is limited. To avoid RRM regulation on these issues, the

scheduler should itself take this into account.

This present proposal mainly consists of a framework that allows for several MAC-scheduling algorithms to obey these constraints without having to explicitly worry for the flow's QoS requirements anymore, as this has been applied by the PDU scheduler already.

In the following an algorithm is used complying with these constraints in a straightforward manner. Later some improvements are shown.

The basic mechanism for MAC scheduling is depicted in Figure 4.

It essentially is the following:

1. Set the queuing pointer in front of the PDU list, i.e. set pointer = 0. Reset the total consumed power $P_{current}=0$.

2. Take the next PDU from the PDU list and consider as much of it in means of TB for scheduling as it is constrained by:

   PDU_size / segment size → result max #1 of TB's

   ARQ-constraint → result max #2 of TB's

       TFC-constrain → result max #3 of TB's

       Maximum single transmission power P_{max}(single): R_{Bi}(max) = P_{max}(single) / C_i (cf. equation 5), where C_i is currently given by equation 2 or equation 3

       → result max #4 of TB's

       Overall power limit P_{limit}: Calculate the virtual available data rate R_{Bi}(available) = (P_{limit} − $P_{current}$) / C_i (cf. equation 4) → result max #5 of TB's

       Formal: Result #TB_{max} = min (max #1 of TB's ... max #5 of TB's);

3. Perform capacity optimizing decisions on TBS creation. It can be advisable to schedule less TB's, with a smaller TBS than possible from the constraints in step 2. If no optimization is desired, chose max #of TB's from step 2. → new variable #Tb$_{schedule}$

5

4. Create #TB$_{schedule}$ TB's. Therefore the segmentation and ARQ on demand is "executed" on them. The such created TB's are assembled to a TBS. Set R$_{Bi}$(used) regarding the created #TB$_{schedule}$.

10

5. Take the TBS and store them together with the associated TF for delivering to PHY-layer in step 8.

6. Compute the total cell power by P$_{new}$ = P$_{current}$ + C$_i$ · R$_{Bi}$(used). Compare this value against a power limit P$_{limit}$.

15

7. If total Power check is ok, i.e. if P$_{limit}$ − P$_{new}$ ≥ P$_{min}$ (P$_{min}$: minimum transmission power for a certain # of TB's) and there are more PDU's in the PDU list, increase P$_{cell}$ by one to the next PDU in the PDU list, set P$_{current}$ = P$_{new}$ and go to step 2.

20

8. Deliver all stored TBS together with their associated TF to PHY layer.

Handling of the power limit P$_{limit}$

This section describes how the power limit P$_{limit}$ for the cell is allocated for MAC-scheduling. The power limit P$_{limit}$ for the scheduler should be chosen according the following rule:

$$P_{limit} = \min\{P_{PS}, P_{current}(t-1) + \Delta P_{inc}\} \qquad \textbf{equation 7}$$

The first term in equation 7 prevents the scheduler to use resources larger than the by RRM assigned P_{PS}. The second term guarantees that the increase in the current transmission power $P_{current}$ is below a given limit $\Delta P_{inc.}$ This limitation is useful in order that the DL power control for all users on other TrCH which are not handled by the scheduler (e.g. users on DCH and of the adjacent cell) can follow the increase of transmission power on downlink shared channel. The overall downlink shared channel transmission power from the previous time $P_{current}(t-1)$, which is estimated by the sum of the transmission powers P_{in}^{prev} of all previous active data flows #i

$$P_{current}(t-1) = \sum_{i \in prev\ active} P_{in}^{prev} \qquad \textbf{equation 8}$$

P_{in}^{prev} (also used in equation 3) is simply the transmission power of the code sequence that was associated to data flow #i (Code Tx power). It can be measured in time periods of some scheduling intervals (up to 100 msec) which is much faster than the overall transmission power measurement (up to 1sec) that is used by RRM.

The basic limit definition in equation 7 can be enhanced as described in the following. For the purpose of efficient handling of the allocated radio resources the MAC Scheduler should monitor the goodput, i.e. the throughput R_{actual} of the scheduler without retransmissions, which simply is defined by

$$R_{actual} = \sum_{i \in active} R_{Bi} \qquad \textbf{equation 9}$$

The virtual bandwidth is defined by the overall available data rate $R_{overall}$, which

can be allocated by the MAC Scheduler. This virtual bandwidth depends on the
allocated transmission power P_{PS} for the scheduler:

$$R_{overall} = function(P_{PS}) \approx P_{PS} / C' \qquad \textbf{equation 10}$$

The value of C' represents a kind of estimate from the constants C_i from all data
flows. The goodput R_{actual} is now compared with the virtual bandwidth $R_{overall}$.
Depending on the comparison result the following actions can be taken:

If $R_{actual} < R_{overall}$, then there is a scheduling problem. The scheduler can process
less data than required. In this case RRM has to be informed to take actions.
This could involve to allocate larger resource P_{PS} to the scheduler if available. If
not, a dynamical resource reallocation to flows has to be performed. This could
for example mean to drop or stop certain flows, that formerly were served with
QoS BW guarantee. Finally this feedback can be used to alter capacity estimates
for future Admission Control decisions.

If $R_{actual} \approx R_{overall}$, then the scheduler works efficient and within the limits. In this
case equation 7 will be used as scheduling policy.

If $R_{actual} >> R_{overall}$, then the scheduler works in a relaxed manner. That means it is
able to schedule much more data than actual required. In this case the scheduler
can have a self limiting behaviour depending on the history of the goodput R_{actual}
in the following way:

if $R_{actual}(t) \leq R_{actual}(t-1)$, then use the following modification of equation 7:

$$P_{limit} = P_{current}(t-1) - \Delta P_{dec} \qquad \textbf{equation 11}$$

where ΔP_{dec} is a certain decrease of transmission power.

if $R_{actual}(t) > R_{actual}(t-1)$, then use equation 7 as it is.

This allows equalisation of the total traffic in terms of traffic shaping. In order to keep the allocated resources available for the scheduling process RRM will not be informed about this self limiting. Nonetheless there will be a noticeable lower variation of power consumption on downlink shared channel which is beneficial for neighbouring cells and DCH power control fluctuation in this cell.

Figure 5 illustrates an example of handling P_{limit}. Both, the usage of the increasing limit ΔP_{inc}, when more power is required, and the usage of the decreasing limit ΔP_{dec} for less required Tx power are shown.

Improvement of MAC-scheduling decision

Especially for NRT services it is not necessary to always attempt to pack the whole PDU into one TBS to be scheduled in one MAC-scheduling interval. It might be desirable to spread the transmission in time over several scheduling intervals. Hence, the invention proposes and uses the following enhancement to section main function:

Main Function

For NRT-services for each PDU the maximum number $N_{schedule}$ of MAC-scheduling intervals $T_{schedule}$ that are allowed for an initial PDU transmission is determined. "Initial" means that this value does not include potential retransmissions. The value $N_{schedule}$ is determined by:

$$N_{schedule} = PDU_{len} / \left(R_{B\min} \cdot T_{schedule} \right) \quad \textbf{equation 12}$$

Here, a system is assumed that obeys no other restrictions (e.g. ARQ..., see assumption mentioned above that if automatic repeat request (ARQ) is applied, the number of retransmissions is assumed to be significantly smaller than the regular traffic). Given this value for each PDU, the MAC scheduler can schedule

5 less TB at a time to reduce the padding. This is possible because the remaining data are apt to fit into a smaller TBS in the next interval.

Conclusion

The invention is not restricted to the specific embodiments described above. A

10 person skilled in the art will recognize that based on a rate conserving strategy the scheduler guarantees the required data rates.

Even if a delay is not explicitly addressed by the scheduling principle, if the respective data flow is in compliance with its required quality of service the

15 scheduler guarantees that there will be no additional delay due to congestion within the scheduling system.

Preferably, bit error rate (BER) requirements are guaranteed in addition via properly forward error correction (FEC) and automatic repeat request (ARQ)

20 functions.

A most preferred application of the improved scheduler is a handling of data flows in the downlink shared channel and the downlink shared channel scheduling was described in detail in the above description. However, the

25 inventive method of QoS scheduling is not limited to downlink shared channels. It also may be applied to scheduling of multiple data flows for different users on dedicated channels (DCHs) in the downlink (DL) and for a single user in the uplink (UL).